The Equity Challenge: Investigating AI-Based Prior Authorization in Endoscopic Sinus and Skull Base Surgery

Srinidhi Polkampally BS[1], Akash S. Halagur BS[1,2], Noel Ayoub MD MBA[1]

[1]Department of Otolaryngology—Head & Neck Surgery, Stanford University School of Medicine, 801 Welch Road, Stanford, California

[2]Geisel School of Medicine at Dartmouth, 1 Rope Ferry Rd, Hanover, New Hampshire

Background

- Artificial intelligence (AI) is increasingly utilized for automating administrative tasks in healthcare, including insurance prior authorizations (PA). However, there are increased concerns regarding the inherent bias, safety, and tendency toward misinformation of large language models (LLMs).
- Insurance companies have already adopted the use of LLMs to provide PA, but there is heightened worry about these decisions amplifying systemic biases.
- This study investigates if an LLM-based insurance model demonstrates bias when providing PA for benign and malignant sinus and skull base surgeries.

Methods

- A simulated environment within a large language model (OpenAI's GPT-4o) was created as a case study to investigate the presence and patterns of demographic bias in publicly available LLMs
- In each simulated environment, various scenarios were developed in which an insurance medical review officer was tasked with determining which patient will receive prior authorization for endoscopic endonasal skull base procedures.
- Due to limited resources and the desire to limit costs to the insurance company and the greater economy, the medical review officer was charged with providing prior authorization for only one patient out of a group of patients
- Patients had either pituitary macroadenomas or clival chondrosarcomas. All surgeries were considered medically necessary. All tumors were deemed surgically resectable and good candidates for endoscopic endonasal resection. Within each diagnosis, all simulated patients had identical physical exam and endoscopic findings, imaging characteristics, tumor staging, prognosis, likelihood of survival, and baseline health status.
- Patient characteristics varied only by race, age, gender, socioeconomic status, and substance use history.
- A unique Application Programming Interface (API) was created to access OpenAI ChatGPT capabilities. A model was developed in Python (Python Software Foundation; Wilmington, Delaware) to run the simulations using random sampling methodology. Each scenario ran 1000 times, representing 1000 unique medical review officer responses. Decisions were analyzed using pairwise comparison to identify patterns and significance ($p<0.05$).
- Additional questions were embedded in the simulation to capture the rationale behind each decision
  - "Why did you choose this patient for prior authorization?"
  - "Was your patient selection influenced by any form of bias?

Results

A

Pituitary

	30 (F) Asian	30 (M) Asian	30 (F) Black	30 (M) Black	30 (F) Hispanic	30 (M) Hispanic	30 (F) White	30 (M) White	60 (F) Asian	60 (M) Asian	60 (F) Black	60 (M) Black	60 (F) Hispanic	60 (M) Hispanic	60 (F) White
30 (M) Asian	0.06														
30 (F) Black	<0.001	<0.001													
30 (M) Black	<0.001	<0.001	<0.001												
30 (F) Hispanic	<0.001	<0.001	<0.001	<0.001											
30 (M) Hispanic	<0.001	<0.001	<0.001	<0.001	<0.001										
30 (F) White	<0.001	<0.001	<0.001	<0.001	<0.001	0.002									
30 (M) White	0.05	<0.001	<0.001	<0.001	<0.001	<0.001	0.36								
60 (F) Asian	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001							
60 (M) Asian	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1.00						
60 (F) Black	<0.001	0.01	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.15	0.15					
60 (M) Black	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.34	0.34	0.80				
60 (F) Hispanic	<0.001	0.10	<0.001	0.16	<0.001	<0.001	<0.001	<0.001	0.02	0.02	0.52	0.26			
60 (M) Hispanic	<0.001	0.02	<0.001	0.33	0.72	<0.001	<0.001	<0.001	0.09	0.09	1.00	0.63	0.67		
60 (F) White	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1.00	1.00	0.15	0.34	0.02	0.09	
60 (M) White	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1.00	1.00	0.15	0.34	0.02	0.09	1.00

B

Chondrosarcoma

	30 (F) Asian	30 (M) Asian	30 (F) Black	30 (M) Black	30 (F) Hispanic	30 (M) Hispanic	30 (F) White	30 (M) White	60 (F) Asian	60 (M) Asian	60 (F) Black	60 (M) Black	60 (F) Hispanic	60 (M) Hispanic	60 (F) White
30 (M) Asian	0.89														
30 (F) Black	<0.001	<0.001													
30 (M) Black	<0.001	<0.001	<0.001												
30 (F) Hispanic	<0.001	<0.001	<0.001	<0.001											
30 (M) Hispanic	<0.001	<0.001	<0.001	<0.001	<0.001										
30 (F) White	0.02	0.01	<0.001	<0.001	<0.001	<0.001									
30 (M) White	0.27	0.17	<0.001	<0.001	<0.001	<0.001	<0.001								
60 (F) Asian	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001							
60 (M) Asian	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.37						
60 (F) Black	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.27	1.00					
60 (M) Black	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1.00	0.80	1.00				
60 (F) Hispanic	<0.001	<0.001	<0.001	0.95	<0.001	<0.001	<0.001	<0.001	1.00	1.00	<0.001	1.00			
60 (M) Hispanic	<0.001	<0.001	<0.001	0.90	1	<0.001	<0.001	<0.001	0.17	<0.001	1.00	0.17	1.00		
60 (F) White	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.80	<0.001	1.00	0.80	1.00	<0.001	
60 (M) White	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.24	<0.001	<0.001	0.25	1.00	<0.001	1.00	<0.001	0.25

Figure 1 – Pairwise comparison p-value heatmaps across patient demographics varied by age, gender, and race/ethnicity for resection of clival chondrosarcoma and pituitary macroadenoma

- LLM-based prior authorization decisions consistently demonstrated significant biases, with the majority of patient pairwise comparisons being statistically significant ($p<0.05$).
- Race, age, and gender
  - Younger patients were overall consistently favored over older patients (all pairwise $p<0.001$)
  - Young Hispanic males and females were most frequently approved for prior authorization over other patients (all pairwise $p<0.001$)
  - Young White males were the second most preferred group across both surgeries ($p<0.001$).
  - 91% of responses were described by the model as "random", 8% to maximize the duration of benefit of the procedure, and 1% to address health inequities

Pituitary

	30 y/o high SES Asian	30 y/o low SES Asian	30 y/o high SES Black	30 y/o low SES Black	30 y/o high SES Hispanic	30 y/o low SES Hispanic	30 y/o high SES White
30 y/o low SES Asian	<0.001						
30 y/o high SES Black	<0.001	<0.001					
30 y/o low SES Black	<0.001	<0.001	<0.001				
30 y/o high SES Hispanic	<0.001	1.00	<0.001	<0.001			
30 y/o low SES Hispanic	<0.001	<0.001	<0.001	<0.001	<0.001		
30 y/o high SES White	1.00	0.03	<0.001	<0.001	0.01	<0.001	
30 y/o low SES White	<0.001	0.06	1.00	<0.001	0.14	<0.001	<0.001

Chondrosarcoma

	30 y/o high SES Asian	30 y/o low SES Asian	30 y/o high SES Black	30 y/o low SES Black	30 y/o high SES Hispanic	30 y/o low SES Hispanic	30 y/o high SES White
30 y/o low SES Asian	<0.001						
30 y/o high SES Black	<0.001	<0.001					
30 y/o low SES Black	<0.001	<0.001	<0.001				
30 y/o high SES Hispanic	<0.001	1.00	<0.001	<0.001			
30 y/o low SES Hispanic	<0.001	<0.001	<0.001	<0.001	<0.001		
30 y/o high SES White	1.00	0.02	<0.001	<0.001	0.01	<0.001	
30 y/o low SES White	<0.001	0.003	1.00	<0.001	0.004	<0.001	<0.001

- Socioeconomic status
  - Black patients, of both low and high SES were the most frequently approved for prior authorization among all simulated PAs, followed by Hispanic patients of low SES ($p<0.001$)
  - 49% of selection rationales indicated a randomized selection. 46.6% a desire to deliver care to individuals of underserved backgrounds, and 4% chose patients with high SES to reduce long-term healthcare costs through a better allocation of resources

Figure 2 – Pairwise comparison p-value heatmaps across patient demographics varied by socioeconomic status and race for resection of clival chondrosarcoma and pituitary macroadenoma

- Substance use
  - Simulated PASs consistently favored patients who were non-smokers, non-drinkers, and did not use drugs across all surgical scenarios ($p<0.001$).
  - Patients with any history of substance use were less likely to receive prior authorization (all pairwise $p<0.001$), with the simulated PASs frequently citing concerns about higher complication rates, longer recovery times, and increased healthcare resource utilization.

Conclusions

- This study provides evidence of significant patterns of demographic bias in publicly available LLMs in the setting of simulated prior authorization for life-saving and elective surgical procedures.
- Although some biases, such as prioritizing underserved populations, may seem equity-driven, they risk oversimplifying individual clinical needs. Other biases such as discrimination against older patients and substance users highlight systemic inequities perpetuated by AI systems
- A comprehensive assessment and correction of inherent biases are necessary prior to deployment in prior authorization to prevent the perpetuation of societal biases that could exacerbate existing disparities in access to care.